Machine Learning 1

Lecture 6.5 - Supervised Learning
Classification - Discriminative Models - The Perceptron

Erik Bekkers

(Bishop 4.1.7)

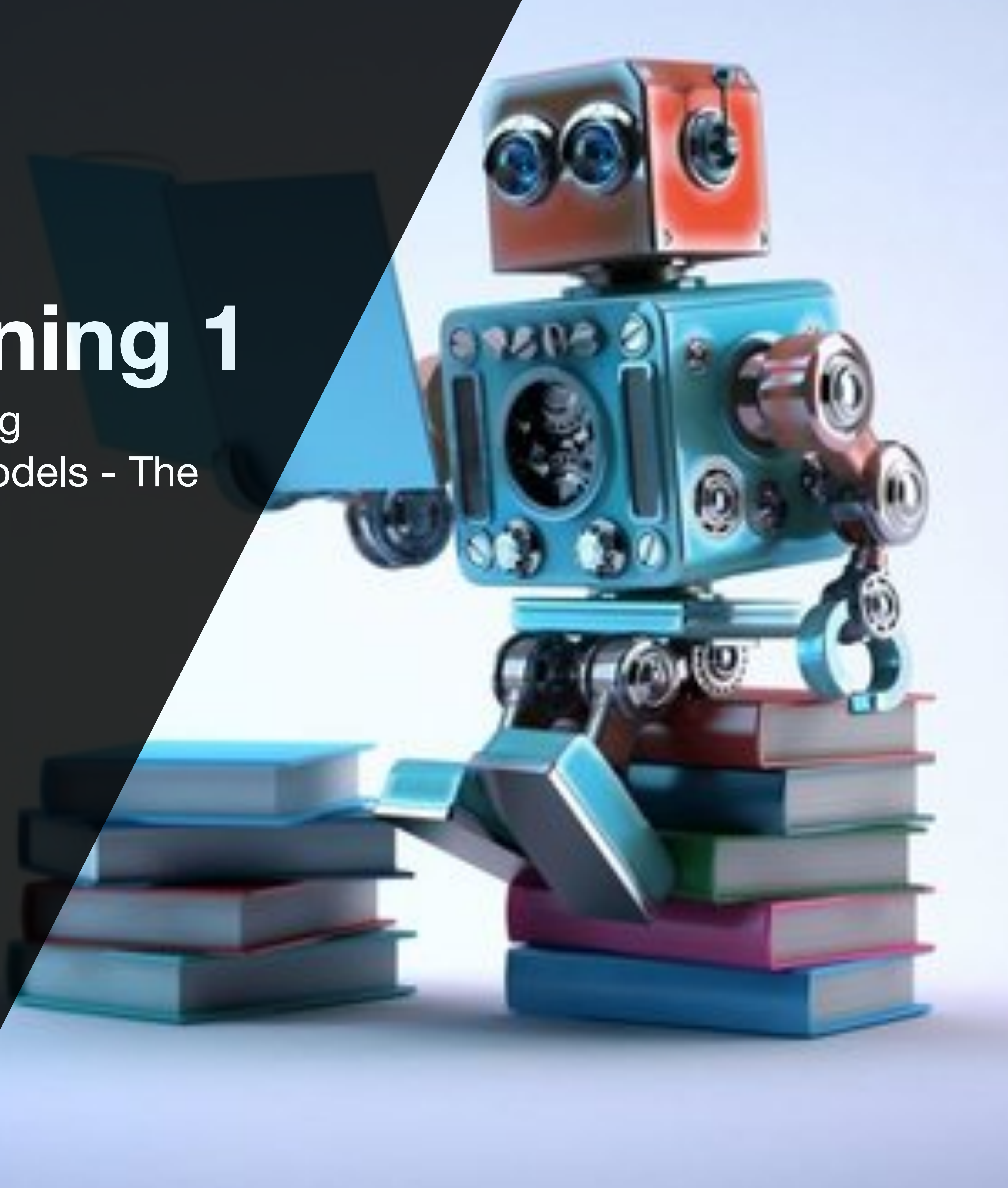
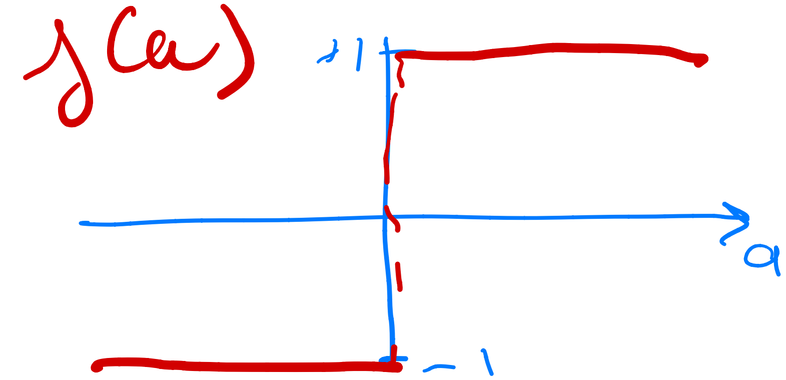


Image credit: Kirillm | Getty Images

Slide credits: Rianne van den Berg

The Perceptron Algorithm



- Input: $\underline{x} \in \mathbb{R}^D$
- targets: $t \in \{C_1, C_2\} \rightarrow t \in \{-1, 1\}$ 2 classes
- Prediction: $y(\mathbf{x}) = f(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}))$ $f(a) = \begin{cases} 1 & , a \geq 0 \\ -1 & , a < 0 \end{cases}$
- Class decisions: assign **x** to class C_1 if $w^T\phi_n > 0$ (and C_{-1} if $\underline{w}^T\underline{\phi}_n < 0$)
- For correct classification: find **w** such that for all (x_n, t_n):

  $\underline{w}^T\underline{\phi}_n t_n \geq 0$

- Perceptron criterion: $E_P(\mathbf{w}) = -\sum_{n \in \mathcal{M}} w^T\boldsymbol{\phi}(\mathbf{x}_n)t_n$ $\leftarrow$ all misclassified

$\mathcal{M}: \{n : \underline{w}^T\phi_n t_n < 0\}$

Perceptron: Stochastic Gradient Descent

✦ $$E_P(\mathbf{w}) = -\sum_{n \in \mathcal{M}} w^T \phi(\mathbf{x}_n) t_n$$

$$= \sum_{n \in \mathcal{M}} E_n(\mathbf{w})$$

$E_n(w) = -w^T \phi_n t_n$

✦ Stochastic Gradient Descent (SGD). For each misclassified $\mathbf{x}_n$:

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_n(\mathbf{w})$$

$$= w^{(\tau)} + \eta(\phi_n t_n)$$

✦If **X** is linearly separable, then perceptron SGD will converge

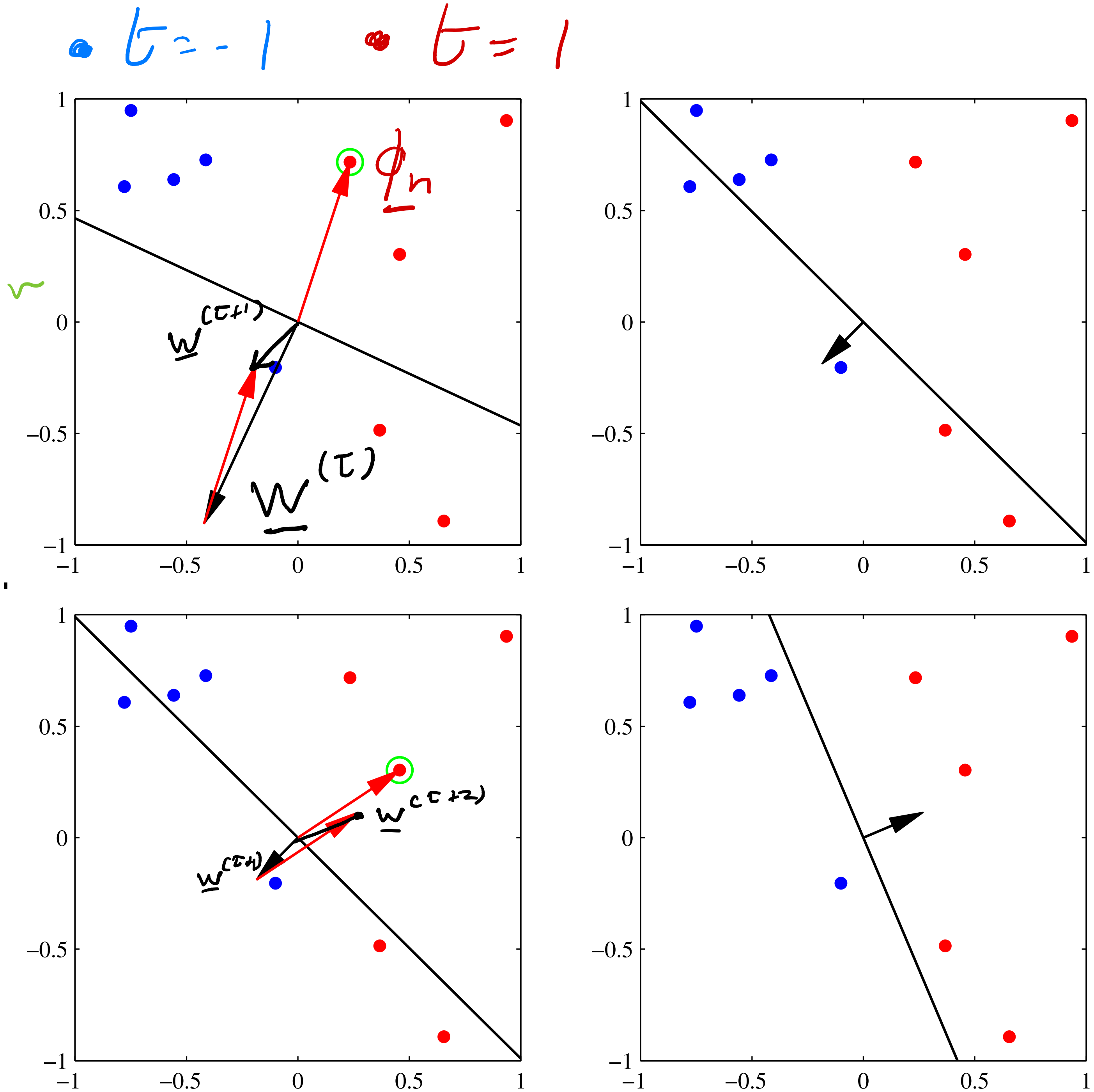


Figure: for $\mathbf{x_n}$ in C_1: add $\phi(\mathbf{x}_n)$ to **w,** for $\mathbf{x_n}$ in C_2: subtract $\phi(\mathbf{x}_n)$from **w.** SGD for perceptron criterion (Bishop 4.7)

Problems: Perceptron

- Perceptron only works for 2 classes

- There might be many solutions depending on the initialization of **w** and on the order in which data is presented in SGD

- If dataset is not linearly separable, the perceptron algorithm will not converge.

- Based on linear combination of fixed basis functions.